Università di Pisa	A.A. 2017-2018

Data Mining II April 10th, 2018

Exercise 1 - Sequential patterns **(16 points)**

A) **(8 points)** Given the following input sequence

<	{A}	{A,B,F}	{B,F}	{E,F}	{G}	{A}	{A,B}	{E} >
	t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7

show all the occurrences (there can be more than one or none, in general) of each of the following subsequences in the input sequence above. Repeat the exercise twice: the first time considering no temporal constraints (left column): the second time considering max-gap = 2 (i.e. gap <= 2, right column). Each occurrence should be represented by its corresponding list of time stamps, e.g..: <0,2,3> = <t=0, t=2, t=3>.

Solutions are highlighted in yellow:

	Occurrences	*Occurrences with max-gap = 2*
***ex.:* <{B}{E}>**	**<1,3> <1,7> <2,3><2,7><6,7>**	**<1,3> <2,3><6,7>**
w_1 = <{A} {B} {E} >	**<0,1,3><0,1,7><0,2,3> <0,2,7><0,6,7> <1,2,3> <1,2,7><1,6,7> <5,6,7>**	**<0,1,3><0,2,3> <1,2,3> <5,6,7>**
w_2 = <{B}{G}{A}>	**<1,4,5><1,4,6> <2,4,5> <2,4,6>**	**<2,4,5> <2,4,6>**
w_3 = <{A,B} {E}>	**<1,3> <1,7> <6,7>**	**<1,3> <6,7>**

B) **(7 points)** For a given dataset of sequences, the GSP algorithm at the **second iteration** found the frequent 3-sequences shown below:

Frequent 3-sequences

{ 2, 4 } → { 5 }	{ 2 } → { 7 } → { 5 }
{ 2, 4 } → { 7 }	{ 4 } → { 5 } → { 5 }
{ 2 } → { 5, 7 }	{ 4 } → { 5 } → { 7 }
{ 4 } → { 5, 7 }	{ 4 } → { 7 } → { 5 }
{ 2 } → { 5 } → { 5 }	{ 7 } → { 5 } → { 5 }
{ 2 } → { 5 } → { 7 }	{ 7 } → { 5 } → { 7 }

Show which candidates the GSP will generate at the **third** iteration, and which of them are removed by pruning.

Answer:

Candidates

1. { 2, 4 } → { 5, 7 }
2. { 2, 4 } → { 5 } → { 5 }
3. { 2, 4 } → { 5 } → { 7 }
4. { 2, 4 } → { 7 } → { 5 }
5. { 2 } → { 7 } → { 5 } → { 5 }
6. **{ 2 } → { 7 } → { 5 } → { 7 }** ← **PRUNED**
7. { 4 } → { 7 } → { 5 } → { 5 }
8. **{ 4 } → { 7 } → { 5 } → { 7 }** ← **PRUNED**

C) (1 point) Which of the candidates generated above are pruned if we are using GSP with a max-gap constraint?

Answer: none of them. Candidates (2) and (3) have no contiguous subsequence, excepted the trivial ones.

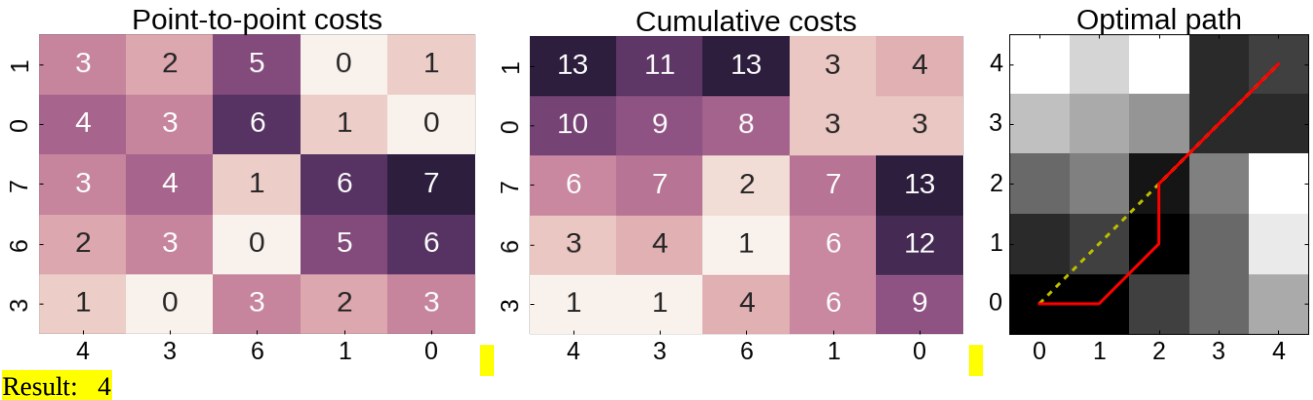
Exercise 2 - Time series / Distances **(16 points)**

A) **(6 points)** Given the following input time series:

t1	< 4, 3, 6, 1, 0 >
t2	< 3, 6, 7, 0, 1 >

compute the distance between "t1" and "t2", using the DTW with distance between points computed as d(x,y) = |x – y|.

Answer:



Result: 4

B) **(2 points)** If we repeat the computation of point (A) above, this time with a Sakoe-Chiba band of size r=1, does the result change? Why?

Answer: No. Because the DTW optimal path remains inside the band.

C) **(1 point)** If we compute DTW(T1,T2), where T1 is equal to t1 in reverse order (namely T1=<0,1,6,3,4>) and similarly for T2 (namely T2=<1,0,7,6,3>), is it true that DTW(T1,T2) = DTW(t1,t2)? Discuss the problem without providing any computation.

Answer: Yes. The optimal path in one direction is the same in the opposite direction. Though, the cumulative costs matrix might look different.

D) **(6 points)** We have to compare a dataset of time series describing the temperature of devices collected every second for a period of one month. In addition we know that they are stationary, and for each time series the values at any time instant are independent from the others and are generated randomly, following approximately a Gaussian distribution. What kind of distance function would you use for any analysis purpose? Please provide as much detail as possible.

Answer: Very long time series => structural distances. Given their properties, a simple set of features for a Gaussian/normal model could be adopted: mean value, variance, kurtosis, skewness.

E) **(1 point)** Computing k-Motifs of width "w" is a complex task. A possible simple way to approximate it, might consist in the following: (i) discretize the input time series into a (long) sequence of symbols; (ii) break the resulting sequence into a set of reasonably long subsequences, for instance to obtain subsequences of length 2*w; (iii) run GSP over the sequences obtained with max-gap=1 and max-span=w; (iv) select the k-most frequent patterns among those of length w.
What kind of errors and limitations will have the results, as compared to the true k-Motifs computation?

Answer: Open question. Points of discussion might include: (a) step (ii) might break time series in the wrong point, losing some patterns; (b) discretization might suffer from the same problem, this time in comparing values (two values of small difference might be put in different classes); (c) k-Motifs use DTW, therefore it is more flexible, again GSP will lose patterns.
